MarineLife-16K: An Attribute-Rich Video-Text Dataset for Motion-Aware Marine Video Understanding

Quang-Trung Truong[1,2], Thanh-Nghia Nguyen[3], Phu-Phong Vo[3], Jaeyeon Kim[2], Cheng Chen[2], Jie Zhang[2], Binh-Son Hua[4], Duc Thanh Nguyen[5], Ivor Tsang[2], Sai-Kit Yeung[1]

[1]Hong Kong University of Science and Technology [2]Centre for Frontier AI Research (CFAR)
[3]University of Information Technology (VNUHCM) [4]Trinity College Dublin [5]Deakin University

Abstract

*Marine video understanding is crucial for a wide range of real-world applications. However, existing marine video datasets suffer from limitations, including the limited number of data samples, fine-grained text descriptions, and detailed visual attributes. Visual attributes—capturing appearance (color, texture) and geometry (shape, size)—comprise a significant portion of scene information, yet they are often confounded by camera motion. To address these limitations, we introduce a large-scale, video-text dataset featuring rich, in-the-wild visual attributes tailored for complex marine environments. Our dataset comprises **2,000** video-text pairs annotated with **13** attributes covering up to four objects per scene, object visual/geometric properties, behaviors, environmental context, and camera motion, spanning diverse lighting conditions, times of day, and 7 countries. From these annotations, we generate **16,080** video-question pairs (**12,080** multiple-choice, **4,000** open-ended). Furthermore, we benchmark both proprietary and open-source video-LLM models on video question answering and video captioning tasks. We find that (i) proprietary models often achieves superior performances; (ii) VLMs handle object motion well but strug-*

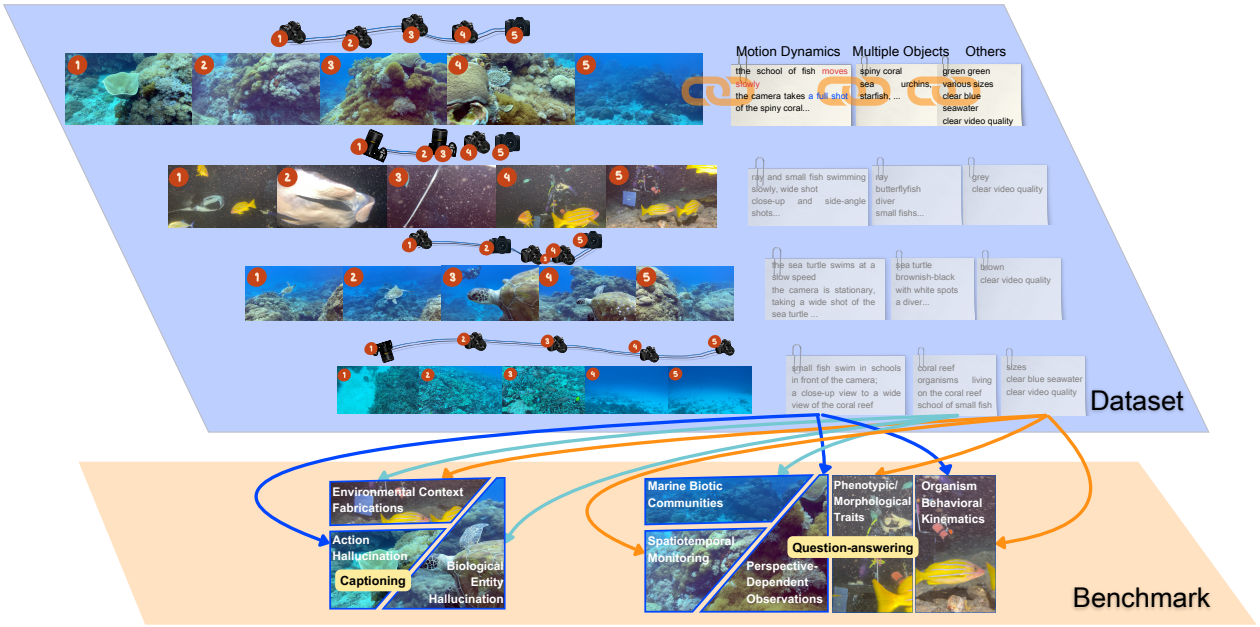


Figure 1. **Annotation Highlights.** We build a dataset of 2,000 video–text pairs, containing fine-grained motion dynamics, including diverse object and camera poses, captured across 7 countries. **Top**: We show video–attribute pairs, highlighting object motion and camera motion to reveal the motion dynamics in marine scenarios. **Bottom**: We benchmark our dataset on video understanding tasks, ranging from captioning to question-answering. We show how Motion Dynamics, Multiple Objects and Others contribute to different dimensions.

Dataset	Time (hours)	Domain	#Sentence	Caption Length	#Video-text Pairs	Cap Source	Attributes			Prompt Complexity
							Object	Motion	Camera	
WebVid [2]	13K	open	-	12	2.5M	generated	✗	✗	✗	low
InternVid-VTT [26]	76K	open	1	10-20	7.1M	generated	✗	✗	✗	medium
MiraData [7]	16K	-	-	-	-	generated	✗	✗	✗	low
MovieBench [29]	69.2	movie	-	-	-	generated	✓	✓	✓	medium
MSR-VTT [31]	41.2	open	1	9.3	200K	manual	✗	✗	✗	low
EPIC-KITCHENS-100 [5]	100	cooking	1	3	19.8K	manual	✗	✗	✗	low
InternVid [27]	760K	open	1	17.6	234M	generated	✗	✓	✗	medium
MovieRecapsQA [19]	107	movie	-	-	60	generated	✓	✓	✗	high
MOMA-QA [4]	147	open	-	-	1412	manual	✓	✓	✗	high
MVK [22]	12.3	underwater	1	9.8	1378	generated	✗	✗	✗	low
MSC [23]	3.6	underwater	2.27	27.4	396	manual	✓	✓	✗	high
MARINELIFE-16K (Ours)	18.48	underwater	3.4	44	2000	manual	✓	✓	✓	high

Table 1. Comparison with existing mainstream video caption datasets. MarineLife-16K features extensive manual annotations for underwater video data, incorporating several attributes, i.e., object awareness, object motion, and camera motion in a hierarchical manner. Note, MovieRecapsQA [19] and MOMA-QA [4] datasets are designed for QA tasks, but our dataset is larger than the QA counterparts.

gle with complex camera motion; (iii) closed-source models struggle to perceive camera motion; and (iv) VLMs struggle to reason in static cameras.

1. Introduction

Marine video analysis presents unique challenges to computer vision, particularly marine life analysis, e.g., fish, whose complex behaviors are intertwined with functional purposes [14, 18]. For instance, schooling behavior serves critical ecological functions, including optimized orientation, hunting, predator confusion, and risk mitigation. Conversely, when individual fish detach from a school to navigate around reefs, their actions often signal a shift toward localized foraging, shelter-seeking, or resting. Capturing these behaviors requires changing from object-centric tracking to motion-centric video understanding [10, 20].

Current video-text benchmarks remain limited when applied to these real-world ecological contexts. Existing datasets focus on global video summarization [34]—describing actions at a broad level without localizing specific timestamps. In contrast, video Question Answering (video-QA) has emerged as a powerful paradigm to evaluate fine-grained reasoning across different temporal windows [32]. However, existing video-QA benchmarks suffer from a lack of motion diversity and fail to challenge models on complex, real-world temporal dynamics.

For QA, it requires datasets that are richly and carefully annotated across different aspects like fact-centric QAs [19], relationship and motion questions [4], focusing on fine-grained attributes. Furthermore, LLM captioning models are often redesigned for QAs, but struggle to perform grounding. To address those limitations, we introduce a novel video-text benchmark dataset for marine video analysis, evaluating dense video captioning and complex motion-aware QA.

Although video-language models (VLMs) are powerful at generating text, they still suffer from hallucinations Fig. 2. Leveraging VLMs for captioning complex videos such as marine footage remains ineffective. In particular, MSC [23] introduces the two-stage video captioning pipeline, aiming to reduce object hallucinations, but fails to detect motion hallucinations in videos. To solve this, we propose a human-in-the-loop captioning engine that reduces motion hallucinations through manual, fine-grained attribute annotations. Compared with MSC [23], our dataset is 5× larger in scale and extracts motion-aware attributes. We compare our MarineLife-16K with existing video-text paired datasets in Tab. 1.

In summary, we make the following contributions in our work.

- We propose a scalable caption engine to reduce hallucinations, using a human-in-the-loop workflow annotated by domain experts.
- We introduce a large-scale motion-driven video dataset with 2,000 video–text pairs, featuring 13 fine-grained attributes for the marine domain.
- We develop a benchmark for question answering (QA) tasks, with 16K question–answer pairs, to evaluate five open-source models and four closed-source models. We observe that VLM models struggle with camera-motion understanding.

2. Related Works

Marine Video Understanding using LLMs. Marine video analysis has recently attracted considerable interest from the computer vision community. Recent works such as MVK [22] and MSC [23] leverage LLMs to understand marine scenarios. More broadly, models like OceanGPT [3], VideoRefer Suite [32], and Video-LLaMA3 [33] integrate spatio-temporal video features and language within a unified embedding space. Specifically, these frameworks map sequential visual tokens into an LLM text space to generate open-ended responses. However, existing LLM architec-

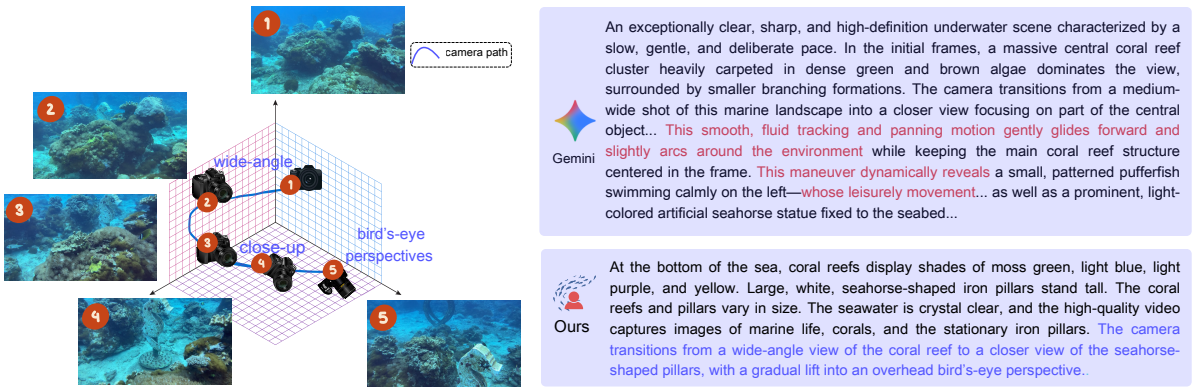


Figure 2. Illustration of motion hallucination. The red texts represent hallucinated camera motion by Gemini 3.1 [6], demonstrating failure to capture camera movement. We capture camera motion ground truths with ❶-❷:wide-angle, ❸-❹:close-up, ❺:bird's-eye perspective.

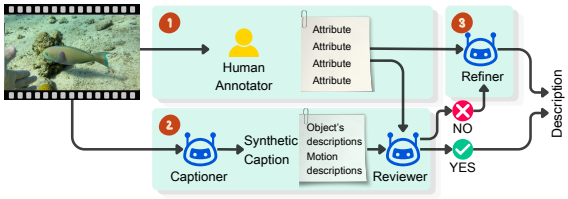


Figure 3. A caption engine. We introduce a three-stage, human-in-the-loop caption engine that utilizes a multi-agent framework to generate high-quality video captions. ❶ Fine-grained attributes are manually annotated by human experts. ❷ A LMM captioner agent inputs video sequences to generate initial synthetic captions, which a reviewer agent subsequently scores against the human-verified attributes ❸ A refiner agent filters the outputs, correcting low-confidence synthetic captions to ensure factual alignment.

tures still struggle with fine-grained temporal comprehension, which can cause hallucinations of complex camera motion and object motion dynamics in challenging underwater scenarios.

Video Question Answering. There has been a growing interest in multimodal VideoQA benchmarks like EgoSchema [12], where models must tightly integrate long-form visual and textual contexts to reason across modalities. Moving beyond generic video understanding, recent benchmarks introduce specialized dimensions of temporal and structural complexity. For instance, HD-EPIC [17] establishes fine-grained, multi-modal egocentric alignments to evaluate action and gaze anticipation over extended timelines, while MovieRecapsQA [19] shifts the focus toward long-form narrative comprehension by grounding open-ended QA pairs in atomic movie facts.

Video-Text Datasets. Recently, many large-scale video–text datasets, e.g., Koala-36M [24], HOIGen-1M [11], and MiraData [7] have leveraged automatic captioning systems (e.g., GPT-4V) to generate video captions. In addition, datasets such as MVK [22] and MSC [23] have recently been introduced for the marine domain, providing insights into VLMs' marine understanding capabilities, including video retrieval and video grounding.

3. MarineLife-16K- A real-world video-text dataset with fine-grained annotations

3.1. Data engine.

We present a scalable, three-stage marine caption engine that leverages human-in-the-loop attributes and multi-agent alignment scoring to generate high-fidelity, factually grounded video descriptions, illustrated in Fig. 3. We use LLM, i.e., GPT-5.5 [15] for dataset construction steps.

- *Step 1: Manual attributes.* Three domain experts extract discrete marine attributes from raw videos.
- *Step 2: Multi-agent-based synthetic caption.* We leverage LLM agents to caption marine videos. Specifically, an LMM Captioner is used to draft text descriptions; an LMM Reviewer scores the semantic alignment between the manual attributes and the synthetic caption, yielding a matching score (S_c), where $S_c \in [0, 5]$.
- *Step3: Final caption.* If $S_c < \tau$, a refiner incorporates human attributes into a synthetic caption to output the final caption; otherwise, it accepts the synthetic caption. We set τ to 3.5 in our work.

3.2. Collection and annotation

Overview. Here, we introduce a real-world dataset, including 2,000 video-text pairs in the marine domain. The example of our MarineLife-16K is illustrated in Fig. 4.

Fine-grained attributes. We also provide fine-grained annotations, including `central object` and up to 3 `additional objects`, `environment`, `lighting`, `video clarity`, `motion/behavior`, and camera attributes, e.g., `angle` and `position` for the collected videos. The captioning annotation begins with manually watching each video entirely to identify a central object and

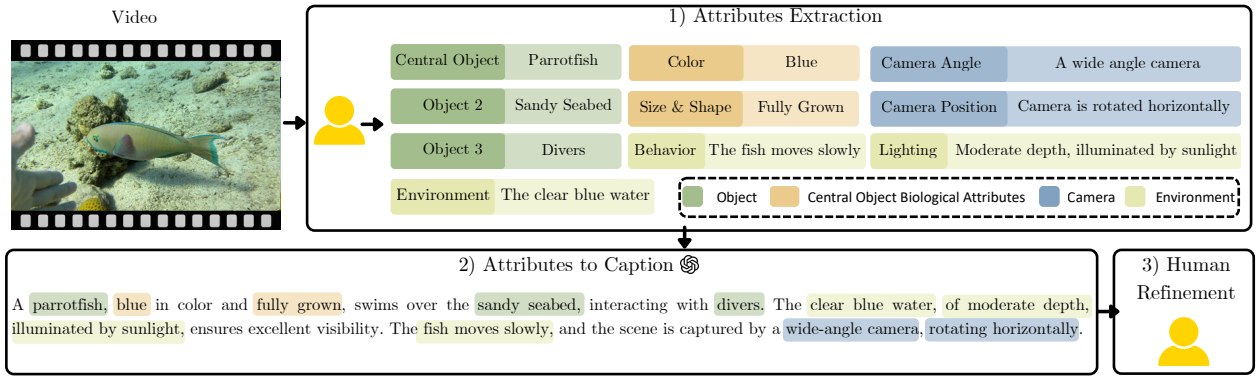


Figure 4. An example of annotated attributes.

any other relevant objects. For each `central object`, we further annotate its appearance attributes, including `texture`, `size`, and `shape&color`. We allow up to 4 objects in each video to be extracted and described in the video's caption. It is important to provide insights of the underwater environment for future ecological studies. To do so, we further describe the environmental factors from the videos, including lighting conditions, video clarity, any movement or behavior observed, and the camera attributes (position, movement, and sharpness), which we describe in detail below. This captioning process takes around 10 minutes per video, excluding the time required to watch the content.

Multi-object. We bound the maximum object density at four objects per clip to maintain descriptive quality. The targeted taxonomy includes common marine entities, such as mobile fauna (fish, human divers), benthic flora (algae, seaweed, seagrass), and foundational substrate organisms (coral and encrusting sponges).

Object motion taxonomy and behavioral categorization. To comprehensively model temporal dynamics, we formalize a fine-grained taxonomy of marine object motions. We decompose behavioral patterns into active kinematics, metabolic functions, and survival strategies. Crucially, our data engine formalizes both the empirical movement speed/behavior and ecological context. We categorize these dynamic visual events into four primary behavioral dimensions: (1) Locomotion and kinematics: tracks fine-grained motion of fins, tails, and body musculature during swimming. (2) Behaviors: encompass fine-grained feeding actions (hunting, stalking, chewing, swallowing, or filter-feeding via suction, biting, and snapping). (3) Survival strategies: models resting/hiding states inside localized substrates (e.g., rocks, coral reefs, vegetation, caves) alongside active defenses such as morphological camouflage, color-changing dynamics. (4) Socio-ecological dynamics: captures coordinated group dynamics (e.g., collective schooling/shoaling) and reproductive events like live birth.

Camera motion dynamics. We describe the dynamic camera using three attributes: *camera position*, *camera angle*, and *video sharpness*. For camera position, we specify whether the camera captures the *full scene, a single viewing angle, or only a specific part of the central object* (e.g., a portion of the main object). For video sharpness, we categorize the footage as either clear or blurry. We also characterize the camera angle and movement. The camera may perform a horizontal pan (turning left/right), a vertical tilt (moving up/down), or a dolly/translation movement in which the camera shifts position while keeping the main object as the reference center.

3.3. Video data quality.

Increasing data diversity. We selected videos, recorded from 27 different locations and at different times, spanning 7 countries. We curated the videos from two datasets, MVK [22] and MSC [23] datasets.

Quality control. We performed quality control by having six human reviewers assess 2,000 text captions.

Fine-grained attributes with motion dynamics. Our dataset contains fine-grained attributes, aiming to reflect the primary challenges of real-world videos, outlined in Table 2.

3.4. Motion-aware QA Generation

For multiple-choice questions, we use fine-grained attributes as answers and use an LLM prompt to generate the remaining answer choices. For open-ended QA questions, we use video captions as facts to create question-answer pairs, evaluating the VLM's ability to understand camera motions. Specifically, the video QA task is designed to evaluate a model's capability to decode the functional relationship between camera movement and the corresponding object's attributes, including motion, appearance, and

Video Challenge	Visual Attribute
Object-Centric	
Heterogeneous Object. Object regions have distinct colors.	Color
Shape Complexity. The object has complex boundaries, such as thin parts and holes.	Size/Shapes
Appearance Change. Noticeable appearance variation, due to illumination changes and relative camera-object rotation.	Lighting , Camera Motion , Size/Shape
Low Resolution. The ratio between the average object bounding-box area and the image area is smaller than 0.1.	Size/Shape
Object & Camera Pose	
Fast-Motion. The average, per-frame object motion, computed as the Euclidean distance between centroids, is larger than 20 pixels.	Behavior
Camera-Shake. Footage displays non-negligible vibrations.	Camera Motion
Motion Blur. The object has fuzzy boundaries due to fast motion.	Lighting , Behavior
Out-of-view. Object is partially clipped by the image boundaries.	Object , Behavior
Occlusion. Object becomes partially or fully occluded.	Object
Scale-Variation. The relative area ratio between any two instances of the target object is restricted to less than 0.5.	Object Motion , Behavior
Interacting Objects. The target object is an ensemble of multiple, spatially-connected objects (e.g., school of fish).	Behavior
Dynamic Background Context	
Dynamic Background. Background regions move or deform.	Background

Table 2. Taxonomy of video challenges and link to visual attributes. We extend attributes [16] to video-text understanding.

geometry. We use GPT-5.5 [15] as LLM judge. Finally, we generate 16,080 video-question pairs, including 12080 multiple choice questions and 4,000 open-ended questions for the collection of 2,000 videos.

3.5. Statistics.

We do statistics on the frequency of text captions of MarineLife-16K. We show frequencies by instance object types (from object 1 to 4), and fine-grained attributes in Fig. 5. In our MarineLife-16K, there are 16,080 questions, including 12080 multiple choice questions and 4,000 open-ended questions for the collection of 2,000 videos.

Dataset Split. We split the dataset into a training set with 1300 videos, a validation set with 200 videos and a testing set with 500 videos.

4. Evaluation Dimensions

To comprehensively evaluate model reasoning ability, we assess VLMs on multiple aspects, including how they summarize overall video content and how well they understand specific questions to test grounding and question-answering (QA) capabilities. We design two benchmarks focused on two key areas: multiple-choice marine-related QA and video captioning in Fig. 7.

4.1. Video-QA

To comprehensively assess vision-language models in the marine domain, we systematically classify our evaluation protocol into five dimensions.

- *Marine Biotic Communities (MBC)* evaluates the model's capacity for multi-object detection and counting. Specifically, we aim to measure the precise identification of the central object as *MBC-so* alongside up to three additional objects as *MBC-mo* to reconstruct the marine biotic community.
- *Phenotypic/Morphological Traits (PMT)* evaluates fine-grained feature extraction for taxonomic classification and biological health assessment. We target the accurate grounding and description of phenotypic traits like texture, size, shape, and color.
- *Organism Behavioral Kinematics (OBK)* assesses VLM's capacity for spatiotemporal motion reasoning from marine video. We evaluate how effectively the model decodes complex motion/behavior patterns.
- *Spatiotemporal Monitoring (SM)* measures VLM models' understanding the environmental context. We assess the integration of environmental attributes, specifically under shifting lighting conditions as *SM-l* and environments as *SM-e*.
- *Perspective-Dependent Observations (PDO)* evaluates the model's viewpoint invariance against dynamic camera pose. Specifically, we test the consistency of marine resource identification across diverse camera angles (horizontal, vertical, centered) and spatial positions (full scene, partial view).
- *Open-Ended with Motion Understanding (OEMU)* assesses how VLMs captures camera viewpoint and objects. We target to object motion, central object and visible activity.

4.2. Video Captioning with Motion Understanding

Our captioning benchmark is used to rigorously measure the model's resistance to hallucination by assigning a dimension score from 0 to 5 to each generated prediction across the following three dimensions.

- *Biological Entity Hallucination (BEH)* evaluates whether the objects referenced in the generated description correctly match the objects specified in the ground truth.
- *Action Hallucination (AH)* assesses whether the described object motion and camera pose are consistent with the actual movements.
- *Environmental Context Fabrications (ECF)* identifies discrepancies by checking whether the generated description includes any facts or elements that are not present in reality.

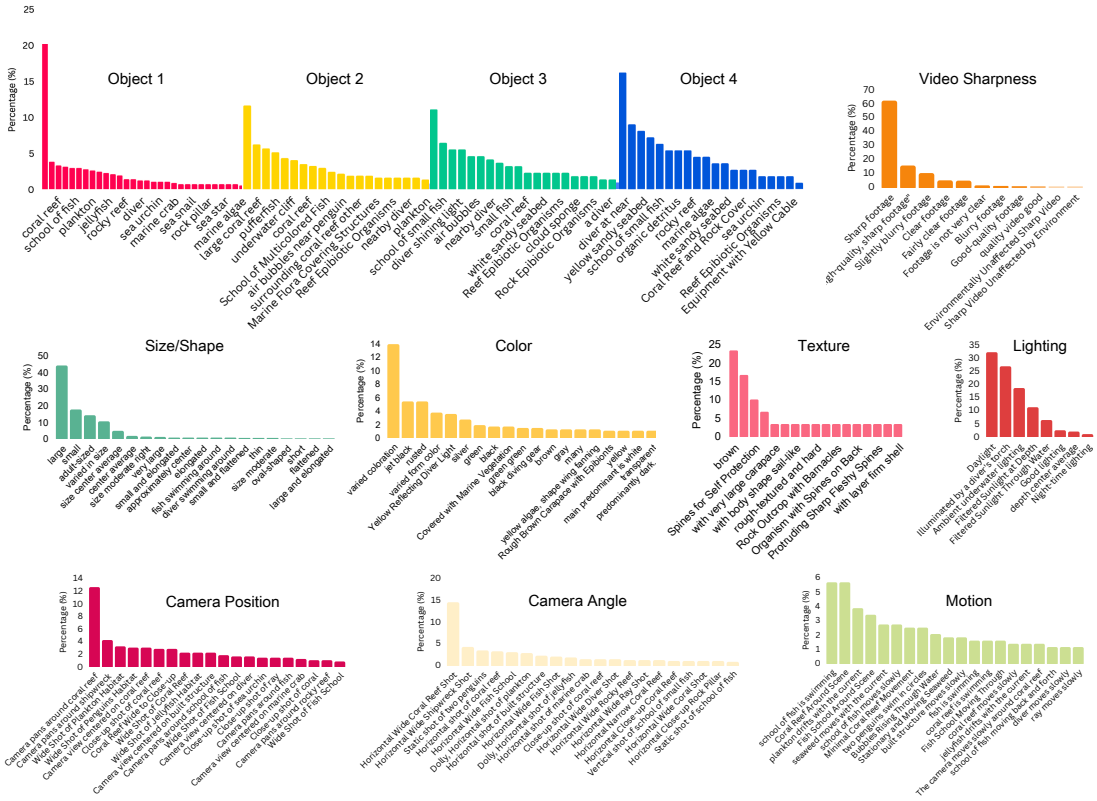


Figure 5. Statistics on the frequency of text-caption of MarineLife-16K. We show frequencies by instance object types (from object 1 to 4) **(Top)**, video quality **(Top Right)**, the size/shape, color, and texture of the central object **(Middle Left)**, camera motion (angle and position) **(Bottom Right)**, and object motion **(Bottom Left)**.

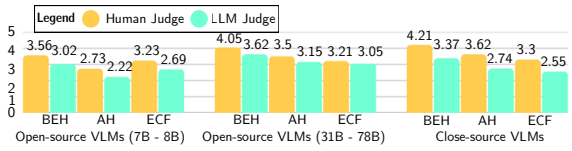


Figure 6. Human study. We recruit *6 annotators* to score LLM predictions across *100* video inputs, requiring *180* total labor hours. We observe both human and LLM judges are consistent when evaluating open-ended QA, which confirms the effectiveness of using LLMs as judges for QA tasks.

5. Experiments

Our goal is to evaluate how state-of-the-art video-LLMs handle VideoQA tasks, focusing on reasoning capabilities that account for motion dynamics, across both open-ended and multiple-choice questions. This section describes our experimental setup and model performance across evaluation dimensions in marine scenarios.

5.1. Models

We evaluate a broad range of proprietary and open-source VLMs for video QA and captioning. Our proprietary lineup includes frontier close-source models, i.e., Claude-4.8-Opus [1], Grok-4.3 [30], Gemini-3.1-Pro [6], and Kimi K2.6 [13]. We further include open-source models, i.e., *small VLMs* (InternVL3-8B [25], LLaVA-OneVision-7B [8], Qwen2.5-VL-7B [28], LLaVA-NeXT-Video [9], Video-LLaMA3 [33]) and *larger VLMs* (Gemma4-31B [21], Qwen2.5-VL-72B [28], InternVL3-78B [25]), using official Huggingface releases ranging from 7B to 78B parameters. We uniformly sample 8 frames per video as an ordered image sequence input. Empirically, we find 8 frames effectively capture camera motion while *balancing* accuracy and token cost. All experiments for the open-source model are conducted using 6 NVIDIA GeForce RTX4090 D. We will release the prompts and hyperparameter configurations used for the public.

5.2. Human Study

To validate the LLM judge for open-ended QA, we conducted a human study with six domain experts, randomly selecting 100 questions across four open-ended QA types evaluating camera motion from visual attributes. Each question was rated by three experts, and scores were averaged. As shown in Fig. 6, despite some gap between human and LLM (GPT-5.5) judges, human scoring is largely consistent with LLM scoring, confirming the effectiveness of using LLMs as judges for QA tasks. Small open-source

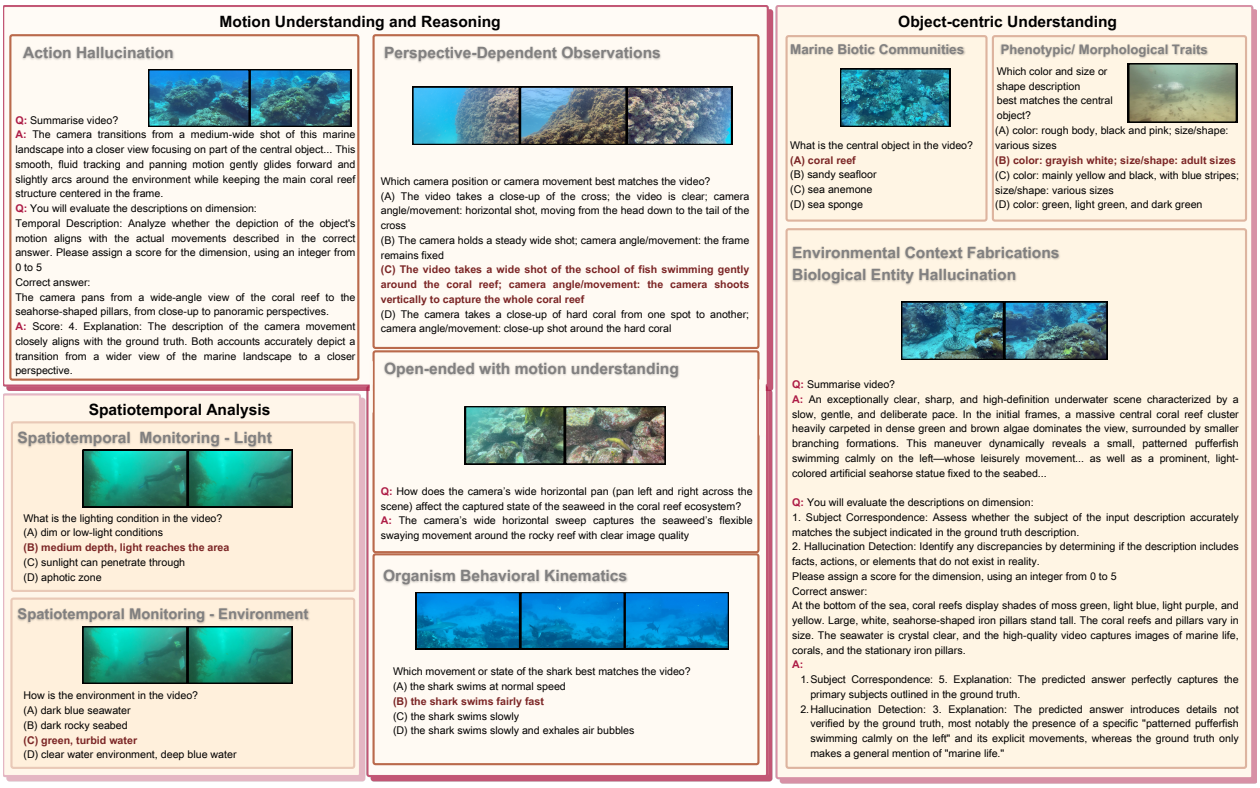


Figure 7. Overview of task dimensions of MarineLife-16K.

Model	Backbone	MBC-so	MBC-mo	PMT	PDO	OBK	SM-e	SM-l	OEMU
Open-source VLMs (7B - 8B)									
Video-Llama3-7B [33]	2D RoPE	96.6	66.6	76.6	63.3	90	73.3	35.3	2.2
LLaVa-NeXT-Video [9]	CLIP	80	56.7	40	60	83.3	23.3	11.8	2.14
LLaVA-OneVision-7B [8]	SigLIP/Qwen	90	63.3	60	53.3	80	83.3	29.4	2.11
Qwen2.5-VL-7B [28]	2D RoPE/Qwen	96.7	66.7	66.7	70	96.7	66.7	29.4	2.43
InternVL3-8B [25]	V2PE/Qwen	96.7	83.3	73.3	70	93.3	73.3	52.9	1.36
Average		90.66	67.32	63.32	**63.32**	88.66	**63.98**	31.76	2.04
Open-source VLMs (31B - 78B)									
Gemma4-31B [21]	SigLIP/Qwen	89.3	36.6	53.3	43.3	86.6	60	47.05	2.52
Qwen2.5-VL-72B [28]	ViT/Qwen2.5	90.5	76.6	80	46.7	93.3	66.7	66.7	2.47
InternVL3-78B [25]	InternViT-6B/Qwen	91.2	63.3	70	46.6	83.4	53.3	29.4	1.98
Average		90.33	58.83	67.77	45.53	87.77	60	47.72	2.32
Close-source VLMs									
Claude-Opus-4.8 [1]	-	93.33	73.3	83.3	60	96.6	60	70.5	2.42
Grok-4.3 [30]	-	93.33	70	76.6	53.3	93.3	60	70.5	2.58
Gemini-3.1-Pro [6]	-	93.3	76.6	90	66.6	96.6	66.6	35.3	2.1
Kimi K2.6 [13]	-	100	60	80	63.3	96.6	63.3	52.9	2.48
Average	-	**94.99**	**69.98**	**82.46**	60.8	**95.78**	62.48	**57.3**	**2.40**

Table 3. Model performance on video question answering. This table reports model performance on our benchmark across MBC, PMT, OBK, SM and PDO metrics. Camera motion-specific metric is challenging for VLM, highlighted in red.

models (7B-8B) struggle with AH.

5.3. Multiple-choice Questions with Motion Understanding

Observation #1: Proprietary models outperform open-source models across most downstream benchmarks. We find that closed-source models outperform open-source models across most evaluation dimensions like MBC-so, MBC-mo, PMT, OBK, SM-l and OEMU in Tab. 3 because closed-source models often train on larger and private datasets rather than open-source datasets.

Model	Static Camera			Moving Camera		
	BEH	AH	ECF	BEH	AH	ECF
Open-source VLMs (7B - 8B)						
Video-Llama3-7B [33]	3.13	2.1	2.6	3.13	2.1	2.43
LLaVa-NeXT-Video [9]	2.9	2.71	2.4	2.43	1.72	2.42
LLaVA-OneVision-7B [8]	3.1	2.29	3.14	3.1	1.85	3.14
Qwen2.5-VL-7B [28]	3.43	2.71	2.71	3.14	2.58	2.57
InternVL3-8B [25]	2.57	1.86	2.3	2.71	1.86	2.42
Average	3.03	2.33	2.63	2.9-0.13	2.02-0.31	2.6-0.03
Open-source VLMs (31B - 78B)						
Gemma4-31B [21]	3.6	3.14	3.02	3.29	3.29	2.85
Qwen2.5-VL-72B [28]	3.57	2.86	2.85	3.71	3.43	3.43
InternVL3-78B [25]	3.29	2.57	2.57	3.57	3.01	3.1
Average	**3.49**	**2.86**	**2.81**	**3.52**+0.03	**3.24**+0.38	**3.13**+0.32
Close-source VLMs						
Claude-4.8-Sonnet [1]	3.14	3	2.3	3.29	2.86	2.43
Grok-4.3 [30]	3.3	2.6	2.28	3.14	2.71	2.57
Gemini-3.1-Pro [6]	3.9	2.4	2.57	3.29	2.57	3.29
Kimi K2.6 [13]	3.6	3.11	2.43	3.14	3.1	2.28
Average	**3.49**	2.78	2.4	3.22-0.27	2.81+0.03	2.64+0.24

Table 4. Comparison of models across static and moving camera settings. Camera motion-specific metric is challenging for VLM, highlighted in red.

Observation #2: While all VLM models exhibit robust capabilities in object motion understanding, they often struggle to capture complex camera motion. The object motion scores OBK of all VLMs are high (average open-source model: 88.66% (7B-8B VLMs), 87.77% (31B-78B VLMs), average close-source model: 95.78%) in Tab. 3. However, the PDO scores are dramatically low (average 7B-8B and 31B-78B open-source models: 63.32% and 45.53%, average close-source model: 60.8%), respectively, and for the close-source group, it is the second lowest across all metrics. This explains why object tracking performs fairly well in which object recognition achieves strong performance with 90.66% and 90.33% for open-source models 7B-8B and 31B-78B, and 94.99% for proprietary models in MBC-so. But when the camera changes, the visual encoder struggles to extract discriminative features, resulting in poor camera motion reasoning in autoregressive language modeling.

Observation #3: The reasoning ability of closed-source models 7B-8B is strong, but limited in camera-motion perception. Although the close-source group dominates across the overall average, in the camera motion PDO metric, open-source models like Qwen2.5-VL-7B (70%) and InternVL3-8B (70.0%) outperforms close-source models like Grok-4.3 (53.3%) and Claude-Opus (60.0%) in Tab. 3. This suggests that Qwen2.5-VL leverages an upgraded Dynamic Resolution ViT, while InternVL3 adopts the Dynamic High-Resolution (DHR) image/frame decomposition strategy. Both enable dynamic token pruning, which results in more effective temporal representation than vanilla average temporal pooling. On the contrary, 31B–78B VLMs often produce detailed captions, but they struggle to predict complex camera motion.

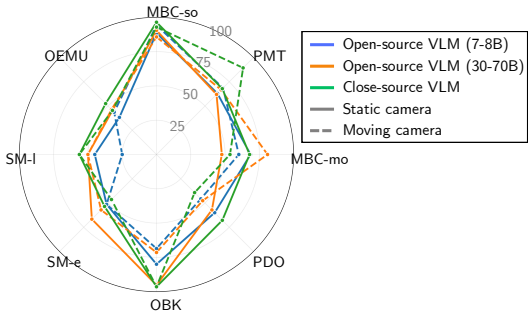


Figure 8. Model performance of VLMs on static and moving camera settings.

5.4. Open-ended Questions

Observation #4: For video summarization, VLMs struggle with camera motion, failing to reason in the static-camera setting. Tab. 4 shows that, for moving cameras, both open-source and closed-source models achieve superior performance compared to static cameras in open-ended QA. VLMs 7B-8B struggle with reasoning capability (ranging from 2.6 to 2.9) across all metrics under camera movement. Thus, we believe VLMs still struggle with camera motion (even static camera), leading to require additional training data with rich camera-motion attributes for fine-tuning. Taking a closer look at multiple-choice questions, Fig. 8 shows VLMs struggle on SM-l, SE-e, and OEMU when the camera is moving, about 50%. In contrast, MBC-so achieves good performance about 90% in both settings. For OBK, performance is slightly higher for moving cameras than for static cameras, yet SM-l remains challenging for VLMs on moving-camera inputs.

6. Conclusion

In this work, we introduced a sequential multi-trajectory expert annotation and refinement strategy, using a human-in-the-loop multi-agent data engine for curating high-quality, motion-centric video-text data for marine environments. Using this strategy, we built a benchmark of 2,000 marine videos with 16,080 fine-grained question-answer pairs, annotated across 13 attributes covering key marine video challenges. We further constructed a comprehensive evaluation suite spanning nine dimensions across open-ended and multiple-choice QA. Our results show both open-source and closed-source models struggle most with camera motion dynamics. We hope this benchmark advances underwater video understanding.

References

[1] Anthropic. Claude 4.8 opus. https://www.

anthropic.com/claude, 2026. Accessed: June 2026. 6, 7, 8

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2

[3] Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. Oceangpt: A large language model for ocean science tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3357–3372, 2024. 2

[4] Wei Dai, Alan Luo, Zane Durante, Debadutta Dash, Arnold Milstein, Kevin Schulman, Ehsan Adeli, and Li Fei-Fei. Towards fine-grained video question answering. *arXiv preprint arXiv:2503.06820*, 2025. 2

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[6] Google DeepMind. Gemini 3.1 pro model card and technical documentation. https://ai.google.dev/gemini-api/docs, 2026. Inference depth control and multi-tier reasoning features. 3, 6, 7, 8

[7] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024. 2, 3

[8] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 7, 8

[9] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 6, 7, 8

[10] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Rushikesh Zawar, Xue Bai, et al. Towards understanding camera motions in any video. *Advances in Neural Information Processing Systems*, 38, 2026. 2

[11] Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24001–24010, 2025. 3

[12] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 3

[13] Moonshot AI. Kimi k2.6. https://www.kimi.com/, 2026. Accessed: June 2026. 6, 7, 8

[14] Peter B Moyle and Joseph J Cech. Fishes: an introduction to ichthyology. 2000. 2

[15] OpenAI. Gpt models. https://platform.openai.com/docs/models, 2026. Accessed: June 2026. 3, 5

[16] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc J. Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 5

[17] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23901–23913, 2025. 3

[18] Tony J Pitcher. *The behaviour of teleost fishes*. Springer Science & Business Media, 2012. 2

[19] Shaden Shaar, Bradon Thymes, Sirawut Chaixanien, Claire Cardie, and Bharath Hariharan. Movierecapsqa: A multimodal open-ended video question-answering benchmark. *arXiv preprint arXiv:2601.02536*, 2026. 2, 3

[20] Xincheng Shuai, Henghui Ding, Zhenyuan Qin, Hao Luo, Xingjun Ma, and Dacheng Tao. Free-form motion control: Controlling the 6d poses of camera and objects in video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12449–12458, 2025. 2

[21] Gemma Team. Gemma 4 technical report, 2026. 6, 7, 8

[22] Quang-Trung Truong, Tuan-Anh Vu, Tan-Sang Ha, Jakub Lokoč, Yue-Him Wong, Ajay Joneja, and Sai-Kit Yeung. Marine video kit: a new marine video dataset for content-based analysis and retrieval. In *International Conference on Multimedia Modeling*, pages 539–550. Springer, 2023. 2, 3, 4

[23] Quang-Trung Truong, Yuk-Kwan Wong, Vo Hoang Kim Tuyen Dang, Rinaldi Gotama, Duc Thanh Nguyen, and Sai-Kit Yeung. Msc: A marine wildlife video dataset with grounded segmentation and clip-level captioning. *arXiv preprint arXiv:2508.04549*, 2025. 2, 3, 4

[24] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025. 3

[25] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6, 7, 8

[26] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[27] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *Inter-*

national Conference on Learning Representations, pages 42055–42079, 2024. 2

[28] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 6, 7, 8

[29] Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28984–28994, 2025. 2

[30] xAI. Grok 4.3. https://x.ai, 2026. Accessed: June 2026. 6, 7, 8

[31] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2

[32] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm-supplementary material. 2

[33] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2, 6, 7, 8

[34] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, 2024. 2